中科院计算所的少数民族语言机器翻译研究进展

吕雅娟 刘群 姜文斌

摘要：本文分析了少数民族语言机器翻译研究的背景、研究现状和发展动态，介绍了中科院计算所在少数民族语言处理和机器翻译方面的研究进展，包括维吾尔语、蒙古语、藏语的语言处理基础技术，形态丰富语言的分析和翻译建模，资源缺乏语言的知识获取和翻译技术，以及组织全国机器翻译研讨会少数民族语言机器翻译评测的情况等。

关键词：少数民族语言；机器翻译；形态丰富语言；机器翻译评测

1 引言

我国是拥有 56 个民族的统一的多民族国家，各民族之间的语言交流还存在严重的障碍。我国的少数民族人口有 1.06 亿，占全国总人口的 8.41%。少数民族语种、文种多，除汉族外的 55 个少数民族中，53 个民族有自己的语言，使用人口 6000 多万；有 22 个少数民族使用着 28 种本民族文字，使用人口约 3000 多万[1]。尽管随着经济社会的不断发展，我国各民族之间的交流越来越频繁，少数民族地区的语言隔阂问题却依然十分严重。教育部的调查数据显示，新疆、西藏仍然有 70%的农牧民不能使用汉语普通话，贵州和云南有 70%～80%的人口不能使用普通话进行交流，新疆仍然有 30%以上的少数民族县、乡镇干部不会说普通话，新疆南部地区 50%～80%的汉语教师的普通话水平在三级甲等以下[2]。与此同时，绝大多数汉族人也不了解少数民族语言。

少数民族语言和汉语的隔阂不仅阻碍了少数民族地区的对外交流和经济发展，而且严重影响了民族团结和社会稳定。一方面，由于语言交流存在障碍，少数民族地区难以及时获取最新的工业、农业、贸易和科技等方面的知识和信息，也难以将独具特色的民族文化向外宣传推广，从而在根本上制约了少数民族地区的发展。另一方面语言隔阂问题给党和国家的方针政策和法律法规在少数民族地区的贯彻执行造成了较大困难，少数民族群众在不明真相的情况下容易受到民族分裂主义势力的蛊惑。进入新世纪以来，境外“藏独”、“疆独”势力一直把少数民族地区作为突破口，加紧利用各种手段鼓吹民族分裂，对国家安全和边疆稳定构成严重威胁。因此，缓解少数民族语言和汉语的隔阂问题对于推动少数民族地区和谐快速发展和维护国家统一具有重要的意义。

机器翻译能够利用计算机将一种语言自动翻译成另外一种语言，是解决这种语言隔阂问题的最有力手段之一。近年来，机器翻译技术特别是统计机器翻译技术取得了巨大的进展，一些语言之间的翻译已经在人们的实际生活和工作中得到了广泛的应用。但是我国少数民族语言和汉语之间的翻译研究进展还比较缓慢，还没有可以实用的系统。相对于目前研究较为成熟的英语和汉语之间的机器翻译来说，我国少数民族语言和汉语之间的机器翻译还面临很多难题和挑战：

– 语言类型跨度大

我国的少数民族语言类型非常丰富。从语言系属分类来看，汉语和藏语同属汉藏语系，但属于不同的语族：汉语属于汉语语族，藏语属于藏缅语族。维吾尔语和哈萨克语属于阿尔

泰语系的突厥语族，蒙古语属于阿尔泰语系的蒙古语族，等等。从语言的形态分类来看，维吾尔语、蒙古语、哈萨克语，朝鲜语等属于形态变化非常丰富的黏着语，而汉语、藏语、彝语、壮语、苗语等属于基本没有词形变化的孤立语。可以看到，各种少数民族语言属性之间的跨度是非常大的，各种语言特征的区别也非常明显，简单地采用现成的经典研究思路在处理如此大跨度的语言翻译时很难取得很好的效果。

– **语言资源缺乏**

现在主流的统计机器翻译方法需要大量的语言资源的支持。如果平行语料库规模不足够大，统计机器翻译效果会受到很大的影响。由于少数民族地区经济文化发展普遍相对落后，可以收集到的语言资源（词典和双语平行语料库等）比汉语少得多。在这种情况下，单纯的统计方法可能很难取得理想的效果，需要融入多种翻译策略和方法，最大程度地利用各种形式的语言学知识和各种资源以提高机器翻译的性能。

– **语言处理基础技术薄弱**

相对汉语来说，一些少数民族语言的处理技术还不够成熟。一些基本的问题，如编码转换、词语切分、词干提取、词性标注、命名实体识别等问题还没有很好解决，而一些更深层次的问题，如句法分析等，还刚刚起步，离在机器翻译中实际应用都还有较大距离，需要进一步深入研究。

这些问题的存在使得目前成熟的一些机器翻译方法对少数民族语言和汉语之间的翻译并不适用。实际上，我国少数民族语言和汉语之间的自动翻译技术面临很多复杂的科学问题，如形态丰富语言的机器翻译，资源缺乏语言的机器翻译等，这些也是目前统计机器翻译研究的重要内容。

2 少数民族语言机器翻译研究现状和发展动态

上世纪九十年代以来，统计机器翻译技术的迅速发展使得机器翻译研究和应用领域都发生了巨大的变化。统计机器翻译的基本原理是为翻译过程构建概率模型，通过对大规模平行文本进行统计分析来估计模型参数，进而使用这些模型参数进行翻译。从 1993 年 IBM 公司首次提出统计机器翻译模型开始，统计机器翻译经历了基于词的模型[3]，基于短语的模型[4, 5]，基于句法的模型[6-9]等几个主要阶段。最近几年，借助语义分析技术改进机器翻译的工作也取得了一定的进展[10, 11]。统计机器翻译技术的发展也大大推动了机器翻译的应用。继谷歌（Google）和微软之后，国内互联网公司百度、网易有道等也相继推出了基于统计技术的在线翻译服务和机译产品。机器翻译在人们日常生活中的应用已经非常普遍。统计机器翻译技术由于克服了传统基于规则的翻译技术中人类专家编写知识所面临的主要困难，而且容易移植到新的领域和语种上，已经成为目前机器翻译学术界和产业界采用的主流技术。

相对于国际上统计机器翻译技术的快速发展，国内少数民族语言机器翻译方面的研究进展比较缓慢，目前研究主要集中在维吾尔语（简称维语）、蒙古语（简称蒙语）、藏语等少数几种语言。

在维语方面，汉维、维汉机器辅助翻译技术的研究起步于上世纪 90 年代中期。1995 年，新疆大学电子工程系王世杰等人的国家自然科学基金项目“新疆民汉语[1]机器翻译系统基础研究”对民汉机器翻译进行过初步尝试[12]。2004 年，新疆大学哈力木拉提等人的新疆自治

1 指“民族语言-汉语”

区科技厅特培专项“计算机汉维辅助翻译软件”为维汉和汉维机器翻译搭建了初步的原型系统[13]。2006 年，新疆大学的国家自然科学基金项目“面向汉维、维汉机器翻译的双语对齐语料库和短语库构建技术的研究”，为汉维双向机器翻译的研究工作做了良好的资源和技术准备。近年来，新疆大学、新疆师范大学等单位都开展了大规模的维汉双语语料库的建设工作[14, 15]，并初步开展了基于统计的维汉翻译方法研究[16, 17]。

在蒙语方面，蒙古语机器翻译经历了探索不同翻译方法的几个阶段。国内学者在汉蒙机器翻译方面曾经做过基于规则的研究[18]和基于实例的研究[19]，并取得一定成果。近年来，也有一些学者在进行基于统计的汉蒙机器翻译的探索[20]；在英蒙、日蒙、蒙汉机器翻译方面也有一些探索性的研究[21-23]。对于蒙古语机器翻译，目前以蒙古语为目标语言的机器翻译研究相对较多，而以蒙古语为源语言的研究则较少。

在藏语方面，自上个世纪 90 年代开始，青海师范大学李延福教授等首次研究汉藏机器翻译技术，先后完成“汉藏科技机器翻译系统”和“汉藏公文机器翻译技术”两项国家“863”计划项目，实现了汉藏科技机器翻译系统和基于规则的汉藏公文机器翻译系统的原型系统。青海师范大学还开展了实用化汉藏机器翻译系统的研究工作；2003 年以来，国内在汉藏机器翻译技术和方法上做了一些理论研究和技术储备工作，包括动词处理、句法分析和命名实体识别[24-26]及藏汉平行语料库的建设[27]。西北民族学院、中国藏学研究中心和中国社会科学院民族研究所等单位在藏语语料库建设以及利用语料库进行藏文信息处理研究方面也都有一些探索和进展[28, 29]。这些储备工作为进一步研究翻译技术奠定了一定的基础。

近年来，少数民族语言和汉语之间翻译研究正得到越来越多的关注与重视。在国家自然科学基金委和科技部的支持下，中科院合肥智能所开展了针对形态丰富语言的统计机器翻译模型构造方法研究，在汉蒙翻译方面取得了较好的效果[30]。北京理工大学开展了基于本体的多策略民汉机器翻译研究，内蒙古师范大学开展了融入语言学知识的汉蒙统计机器翻译研究，等等。

2011 年第七届全国机器翻译研讨会（China Workshop on Machine Translation, CWMT）[2]机器翻译评测首次引入了少数民族语言到汉语的翻译评测项目，进行了包括维语、蒙语、藏语、哈萨克语以及柯尔克孜语 5 种民族语言到汉语的翻译评测任务，共有 10 家研究机构和大学参加了该次评测[31]。在评测中我们发现，参加民族语言翻译评测项目的来自 10 家单位的 24 个系统，包括少数民族院校提交的系统，全部采用了基于统计的翻译技术。可以看出统计机器翻译技术已经在少数民族语言机器翻译研究中得到了广泛重视。但是，从少数民族语言和汉语间翻译的特点和少数民族语言处理研究的现状看，直接应用现有的统计技术还存在很多问题。首先，民族语言和汉语间的语言类型差别大，用同样的模型解决所有的语言对之间的翻译问题是行不通的。主流的统计翻译模型将任何一种语言都同等对待，对于形态差异较大的语言对（如维语、蒙语等黏着语和汉语），直接利用现有的统计机器翻译模型并不能很好地描述语言对间的差异，翻译结果也不理想。其次，民汉翻译资源非常缺乏，单纯的统计方法并不能得到很好的翻译效果。事实上规则方法和统计方法各有优缺点。规则方法更容易有效利用专家知识，对于比较规律的语言现象，如时间词、数词等也可以实现高精度的翻译。在资源缺乏的情况下应该考虑综合利用规则和统计等多种翻译策略。此外，一些少数民族语言处理基础技术还很薄弱，缺乏高性能的词法分析、命名实体识别等工具，句法分析研究目前还都很不成熟，这些基础技术对于翻译模型的选择，以及翻译模型的训练等都会产生很大的影响。

[2] http://mt.xmu.edu.cn/cwmt2011/

通过以上分析，我们认为少数民族语言机器翻译研究在充分借鉴现有统计机器翻译研究方法和经验的基础上，还应该更加注意结合语言本身的特点。一方面，需要进一步加强少数民族语言处理基础技术研究，以更好地支持机器翻译等应用和系统开发；另一方面，需要研究适合少数民族语言和汉语的翻译模型和方法，如形态丰富语言的机器翻译方法、面向资源缺乏语言的机器翻译方法等。这些问题的深入研究对于解决很多小语种的机器翻译问题，进一步推动机器翻译研究的发展都具有重要的意义。

3　中科院计算所的少数民族语言处理和机器翻译研究进展

我们中科院计算所自然语言处理研究组专注于机器翻译研究 20 余年，曾经开发过基于规则、基于实例和基于统计的机器翻译系统。近十年来，研究组在基于统计的机器翻译研究和应用方面取得了较大的进展。研究组提出了一系列基于源语言句法分析的统计翻译模型，在本领域最有影响的国际期刊（Computational Linguistics）和学术会议（ACL[3]，EMNLP[4]，COLING[5]）上发表相关论文 50 余篇，申请技术发明专利 18 项，受到国内外同行的广泛关注和跟踪。研究组开发的机器翻译系统在著名国际机器翻译评测 NIST[6]和 IWSLT[7]中多次取得好成绩。研究组还将统计机器翻译技术实际应用到了专利翻译、移动翻译、新闻翻译等多个领域中。

近年来，我们组开展了少数民族语言和周边国家语言的机器翻译研究。在少数民族语言方面，我们主要关注维吾尔语、蒙古语、藏语等我国使用人口最多的几种少数民族语言。我们与新疆大学、内蒙古大学和青海师范大学建立了紧密的合作关系。经过几年的努力，在维吾尔语、蒙古语、藏语处理以及它们和汉语间的机器翻译方面取得了较大的进展。我们收集加工了较大规模的维汉、蒙汉、藏汉平行语料库和翻译词典，开发了一系列初步实用的民族语言处理基础工具，如语种识别和编码转换工具、维语形态分析工具、蒙语形态分析工具、藏语断句/分词工具、命名实体识别和翻译工具等，研究了面向形态丰富语言的翻译模型和资源缺乏语言的翻译方法，搭建了维汉、蒙汉和藏汉统计机器翻译系统。我们开发的少数民族语言翻译系统已经在国家有关部门得到了应用。研究组还负责组织了全国机器翻译研讨会少数民族语言机器翻译评测，为推动国内少数民族语言机器翻译的发展做出了贡献。

本节将介绍我们在少数民族语言处理和翻译研究方面的主要进展，下一节将介绍我们组织全国机器翻译研讨会少数民族语言机器翻译评测的情况。

3.1 维、蒙、藏语言处理基础技术

语言处理是机器翻译的基础。无论是对于机器翻译本身，还是机器翻译所需要的语料库处理来说，基本的语言处理技术都是不可或缺的。对于维、蒙、藏语的处理，我们重点解决了语言编码、形态分析、分词和命名实体的识别等机器翻译所必需的基本语言处理技术。下面分别进行简单的介绍。

– 编码识别和转换

维语、蒙语、藏语等很多少数民族语言都存在多种编码形式，如藏语常用的编码除了 Unicode 外，还有班智达、华光、同源、桑布扎码等等，为了对这些语言进行处理，必须首

[3] Annual Meeting of the Association for Computational Linguistics
[4] International Conference on Empirical Methods in Natural Language Processing
[5] International Conference on Computational Linguistics
[6] NIST Open Machine Translation (OpenMT) Evaluation: http://www.itl.nist.gov/iad/mig/tests/mt/
[7] The International Workshop on Spoken Language Translation (IWSLT): http://iwslt.org/

先进行编码的识别和转换。为了同时支持多种语言文本的处理，还要进行语种的识别。由于语种和编码在计算机内部表示上可以统一看成是编码问题，我们把语种识别和编码识别问题同时考虑，采用统一的模型和方法进行处理。我们提出了一种通用的基于统计语言模型的语种和编码识别方法[32]。首先将编码粗识别为三类字符编码系列，然后结合三种粒度语言模型同时实现语种和编码的识别。该方法不依赖于各种少数民族语言特有的规则，便于扩展到新的语种和编码。系统中的三种粒度语言模型分别是基于字节的语言模型、基于字符的语言模型以及基于词的语言模型。三种粒度的语言模型分别从三个层面区分语种和编码，能够更好地完成识别任务。系统总的处理流程如图 1 所示。

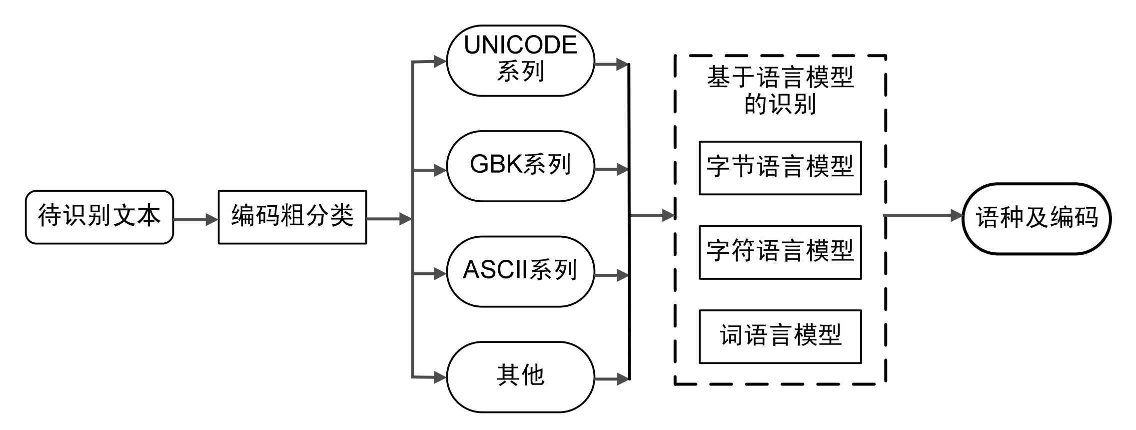


图1. 多语种文本语种和编码识别系统流程图

基于该方法我们实现了一个语种和编码识别工具，目前支持 11 种语言及其主流编码，包括：汉语、英语、藏语、维语、蒙语、阿拉伯语、土耳其语、俄语、哈萨克语、柯尔克孜语、日语，识别的平均准确率大于 95%。对于藏语、维语、蒙语主流编码的转换工具也已经完成。

– **维语、蒙语形态分析**

维语、蒙语都属于形态丰富的黏着语，词语通常由词干和若干词缀组成，形态分析的任务就是解析出词语的词干和词缀结构，并且标定出它们的类别。针对黏着语的构词特点，我们设计了一种基于图状结构的判别式模型。该模型将句子的形态分析结果表示为图状结构，并通过特征设计，以图中的边描述词语内部形态成分之间以及分属相邻词语的形态成分之间的关联约束。具体而言，在图状模型中，每个词语内部各词干和词缀之间都存在相应的边，对应相应的近距离特征；分属相邻词语的词干和词缀之间也存在相应的边，对应相应的远距离特征。这两类特征分别描述了词语内部和词语之间各形态成分之间的语言学关联约束关系。与传统的线性模型、图状模型相比，该模型更好地考虑了各形态成分之间的语言学关联。实验表明，基于图状建模的形态分析与线性建模方式相比，取得了显著的性能提升，并且显著超越了前人的相关工作。

基于图状模型，我们已经实现了初步实用的维语、蒙语、韩语形态分析工具。本专辑将另有文章详细介绍该工作的最新进展。相关工作也可以参见我们已经发表的论文[33, 34]。

– **藏语分词**

藏语的构词模式比汉语要复杂得多。我们根据藏文的构字和构词特性，有机结合规则方法和统计方法的优点，构建了适合藏文的词语切分模型。首先，根据藏文特有的构词规律将句子切分成最小粒度的序列，称之为单元序列；然后，根据感知机模型提供的判别式分类的权重，在单元序列上进行粗切分，从而生成有向图，并通过查询词典为有向图的边赋予不同

的权重；最后，通过动态规划算法求解加权有向图中的最短路径，生成最终分词结果。图 2 给出了藏文分词系统的工作流程。

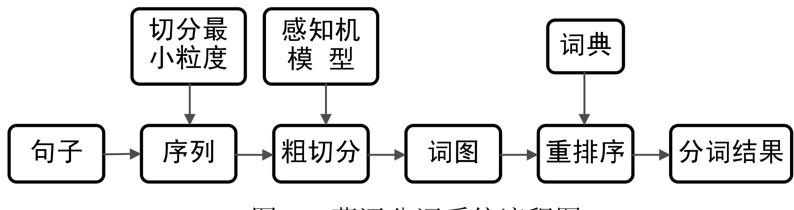


图2. 藏语分词系统流程图

实验证明，基于判别式模型和词图重排序技术的藏文词语切分模型较之前人最好的工作有了显著进步[35]。基于该方法实现的藏文分词工具已经实际用于我们的藏汉机器翻译系统中，取得了很好的效果。本专辑将另有文章对该工作进行详细的介绍。

– 命名实体识别和翻译

命名实体的识别是语言分析的重要环节。时间词、数词、人名、地名和机构名等命名实体的正确识别，对自然语言处理后续阶段，如句法分析和机器翻译都大有助益。尽管汉语和英语命名实体识别技术已经比较成熟，但是由于维语、蒙语、藏语自身复杂的语言特点，不能简单套用现成的理论模型和方法。比如说，维语的命名实体可缀接复杂的后缀，这使得维语的命名实体识别和形态分析任务密不可分，必须进行特殊的处理。对于时间词和数词的识别和翻译，由于各个语种都具有较强的规律性，我们采用了人工语言学规则加双语词典的模式进行处理。对于人名、地名、机构名等实体的识别和翻译，我们希望能够建立一个通用的多语种命名实体识别和翻译框架。在核心算法上采用与具体语言无关的统计方法，同时结合规则方法对具体的语言现象进行专门处理。我们提出了规则知识和统计建模相结合的命名实体识别系统框架，以期既充分利用统计模型稳定性好、精度高和语言无关的优势，又能充分考虑各个语种特有的词法和句法规律，以取得更好的命名实体识别精度。

目前我们已经实现了维语、蒙语、藏语的数词、时间词识别，构建了翻译工具，达到了较高的识别和翻译准确率，有效提高了翻译质量。基于判别式统计模型的命名实体识别引擎也已经开发完成，还需进一步扩大维语、蒙语和藏语命名实体标注语料库，以及命名实体翻译词典的规模。

3.2 形态丰富语言的分析和翻译建模

形态丰富语言包括黏着语（如芬兰语、日语、韩语等）和部分形态变化比较复杂的屈折语（如德语、法语、阿拉伯语、俄语等）。我国的很多少数民族语言，如维吾尔语、蒙古语、哈萨克语、朝鲜语等都属于形态丰富语言。形态丰富语言每个词的变化形式最多可达数百种，甚至上千种。而目前机器翻译研究界关注最多的汉语和英语都属于形态变化比较简单的语言。汉语基本上没有形态变化，英语形态最丰富的动词也只有四、五种变化形式。现有主流的机器翻译方法基本上不考虑词形变化，把每个不同词形的词都当成独立的词语来考虑。但是对于形态丰富的语言，这种做法就会带来比较严重的数据稀疏问题，会导致翻译时出现大量的未登录词，严重影响机器翻译的性能。除了形态变化丰富以外，形态丰富语言中的很多句法特性（如时态、语态、人称、数等）也都是通过动词的形态来表达的，而在形态简单的汉语或英语中，这些句法特性大部分都通过特定的词语来表达。这就导致这两类语言的句法同构性非常差，而现有的机器翻译模型对于这种结构差异较大的语言之间的翻译效果都不理想。

为了能够很好地实现形态丰富语言和简单形态语言之间的机器翻译，必须在语言的更深层次上实现两种语言的映射，使得翻译模型能够充分把握形态丰富语言的特性及其与汉语之间的互译对应规律。

目前常见的词语表示形式有两种：一种是把完整的词表示成一个独立的单位。由于形态丰富语言中有些词的变化形式有数百甚至上千种之多，这种词语表示形式会在机器翻译中造成出现大量未登录词的严重问题。另外一种表示方法是将完整的词切分成词干加多个词缀形式，每个词干和词缀作为一个独立的单位。这样虽然减轻了未登录词问题，但会使词干之间距离变得比较远，大大削弱统计模型的有效性。其实，词语的词干、词缀表示形式是可以非常灵活的。例如图 3 中的多粒度的线性词语表示形式，以及图 4 中基于图的词语表示形式等。图中 A 表示词干，B 表示词缀。

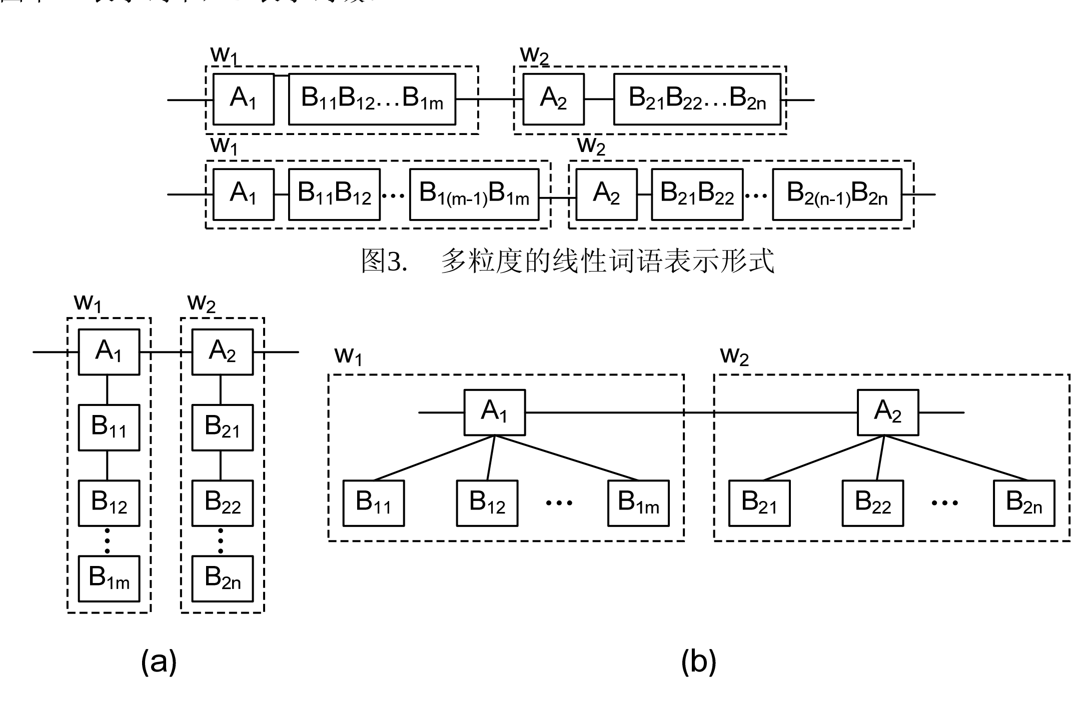


图3. 多粒度的线性词语表示形式

图4. 基于图的词语表示形式

不同的词语表示形式对于词法分析、词语对齐、机器翻译的建模和算法都有一定的影响。我们希望对各种形式的词语表示形式进行深入的研究，以探讨机器翻译最合理的表示形式以及相应的词语对齐、翻译模型和算法。基于图 4（a）的图状词语表示形式，我们实现了基于图状结构的形态丰富语言形态分析模型，在维语、蒙语和韩语的形态分析上取得了很好的分析效果[33, 34]。在形态丰富语言的翻译建模方面，我们通过多种粒度表示，区别对待词干、词缀等方式来改善形态丰富语言到汉语的翻译质量，在维吾尔语、哈萨克语和柯尔克孜语上都取得了显著的效果[36, 37]。本专辑将另有文章详细介绍这方面的工作。

3.3 资源缺乏语言的知识获取和翻译技术

语言资源缺乏是构建少数民族语言到汉语翻译系统所面临的主要问题。目前的统计机器翻译方法是建立在大规模双语平行语料库的基础上。如果没有大规模双语语料库，统计机器翻译方法的优势，如开发成本低、周期短等，就都不存在了。但是对于大部分语言对来说，大规模的双语平行语料库的获取并不容易。我国少数民族语言资源建设虽然近年来得到广泛关注和发展，但是可以收集到的双语平行语料资源都很有限。另一方面，我国少数民族民族语言的自然语言处理基础相对比较薄弱，缺乏既熟悉少数民族语言又精通基于规则的机器翻译方法的专家，基于规则的翻译系统更难实现和维护。因此，对于资源缺乏的少数民族语言

来说，单纯的统计方法或规则方法都可能很难取得理想的效果。最大程度地利用各种形式的语言资源和人力资源实现快速的知识和资源获取，同时有效融合多种翻译策略以提高机器翻译系统的性能，是解决资源缺乏语言的知识获取和翻译问题的有效途径。

针对语言资源缺乏问题，我们希望通过人机交互的方式，确定现有的机器翻译系统的知识盲点，有针对性地引入人类专家的经验知识，与机器自动学习的过程紧密结合，从而加快学习的进度，更有效地综合利用人类专家知识与机器统计学习的能力，来改善机器翻译的效果。我们已经尝试将人类专家撰写的规则融入到统计翻译系统中，用以解决长距离调序和句子骨干翻译问题，取得了较好的效果[38]。此外，我们对多粒度融合的词汇对齐策略[39]，基于双语映射的无监督少数民族语言句法分析知识获取策略[40]等进行了研究。这些工作部分缓解了语言资源缺乏的困难，取得了一定的效果。语言资源缺乏的另一种情况是领域资源的缺乏。应对这种情况需要解决机器模型的领域自适应问题。这方面我们也已经开展了一些工作，但是目前效果还不够理想。

基于以上研究，我们开发了一系列初步实用的少数民族语言处理基础工具，搭建了维汉、蒙汉和藏汉统计机器翻译系统。目前我们的少数民族语言翻译系统已经在国家有关部门得到了应用，得到了用户的好评。除少数民族语言外，研究组还实现了韩语、日语、泰语、俄语、阿拉伯语、越南语到汉语的机器翻译原型系统，在韩语形态分析、日语分词，泰语分词方面也取得了一定的进展。

4 少数民族语言机器翻译评测

全国机器翻译研讨会（China Workshop on Machine Translation, CWMT）由中科院自动化所、计算所和厦门大学于 2005 年联合发起[8]，旨在推动中国机器翻译研究的发展，促进国内外同行的交流。研讨会从 2007 年开始举办机器翻译评测活动（简称 CWMT 机器翻译评测），目的是更加有效地推进研究单位间实质性的交流、促进机器翻译技术的发展。中科院计算所自然语言处理研究组负责了机器翻译评测活动的组织工作。2011 年，在研究组的倡导下，第七届全国机器翻译研讨会（CWMT 2011）机器翻译评测首次引入了少数民族语言到汉语的翻译评测项目，进行了包括维语、蒙语、藏语、哈萨克语以及柯尔克孜语五种民族语言到汉语的翻译评测任务。2013 年即将举办的第九届全国机器翻译研讨会[9]（CWMT 2013）机器翻译评测将继续举办维汉、蒙汉、藏汉三个民族语言项目的评测。表 1 和表 2 分别给出了这两次机器翻译评测的项目设置情况。其中灰色背景的为少数民族语言翻译评测项目。评测组织方中科院计算所联合各少数民族院校为参评单位提供了训练语料。表 3 给出了这两次评测中少数民族语言评测项目提供的训练语料规模，以及语料提供单位。

CWMT 2011 和 CWMT 2013 少数民族语言机器翻译评测吸引了包括中科院计算所、自动化所、新疆理化所、哈尔滨工业大学、东北大学等共 14 家单位参加。CWMT 2013 评测中我们还联合中科院自动化所、厦门大学分别为蒙汉、维汉、藏汉评测项目提供基线系统(Baseline)，包括训练、解码全过程的源码和相关工具。少数民族语言机器翻译评测给国内从事机器翻译的研究单位和从事少数民族语言信息处理的单位提供了合作和交流平台，这必将进一步促进少数民族语言机器翻译研究和应用水平的提高。我们期待着更多的研究团队能参加到这个评测中来。

[8] 最初三届研讨会的名称为“全国统计机器翻译研讨会”，从 2008 年起更名为“全国机器翻译研讨会”
[9] http://www.liip.cn/CWMT2013/

表1. CWMT 2011机器翻译评测项目

序号	评测项目名称	语种	领域
1	汉英新闻领域机器翻译	汉语→英语	新闻领域
2	英汉新闻领域机器翻译	英语→汉语	新闻领域
3	英汉科技领域机器翻译	英语→汉语	科技领域
4	日汉新闻领域机器翻译	日语→汉语	新闻领域
5	蒙汉日常用语机器翻译	蒙语→汉语	日常用语
6	藏汉政府文献机器翻译	藏语→汉语	政府文献
7	维汉新闻领域机器翻译	维语→汉语	新闻领域
8	哈汉新闻领域机器翻译	哈萨克语→汉语	新闻领域
9	柯汉新闻领域机器翻译	柯尔克孜语→汉语	新闻领域

表2. CWMT 2013 机器翻译评测项目

序号	评测项目名称	语种	领域
1	汉英新闻领域机器翻译	汉语→英语	新闻领域
2	英汉新闻领域机器翻译	英语→汉语	新闻领域
3	英汉科技领域机器翻译	英语→汉语	科技领域
4	蒙汉日常用语机器翻译	蒙古语→汉语	日常用语
5	藏汉政府文献机器翻译	藏语→汉语	政府文献
6	维汉新闻领域机器翻译	维吾尔语→汉语	新闻领域

表3. 少数民族语言评测项目训练语料情况

评测项目	CWMT 2011	CWMT 2013	语料提供单位
维吾尔语 → 汉语	5 万句对 (1,091,903 维语词)	11 万句对 (1,912,542 维语词)	新疆大学 中科院新疆理化所
蒙古语 → 汉语	6 万句对 (982,135 蒙语词)	10.7 万句对 (2,251,117 蒙语词)	内蒙古大学 中科院合肥智能所
藏语 → 汉语	10 万句对 (1,280,837 藏语词)	12.6 万句对 (1,391,752 藏语词)	青海师范大学，厦门大学， 西北民族大学，西藏大学
哈萨克语 → 汉语	5 万句对 (965,570 哈语词)	——	新疆大学
柯尔克孜语 → 汉语	5 万句对 (1,175,823 柯语词)	——	新疆大学

5 总结与展望

近几年来中科院计算所自然语言处理研究组在少数民族语言处理和机器翻译方面做了不少工作，取得了一定进展。本文对相关工作做了一个概况性的说明，一些细节在本专辑的其他文章中有更详细的介绍。今后，我们将进一步扩大少数民族语言翻译资源的规模，研究适合少数民族语言和汉语间机器翻译的关键技术和方法，希望能够显著提高维吾尔语、蒙古语、藏语等主要少数民族语言与汉语之间的自动翻译水平，推进少数民族语言处理技术的发展和机器翻译系统的实用化。我们也将继续开展少数民族语言机器翻译评测活动，希望少数

民族语言处理和机器翻译研究和应用能够得到越来越多的关注与重视。

参考文献：

[1] 《关于进一步繁荣发展少数民族文化事业的若干意见》学习辅导读本. 2009. 国家民族事务委员会编发。http://cpc.people.com.cn/GB/165240/167240/10085487.html

[2] 语言障碍影响少数民族学生升学就业. 中国青年报. 2005 年 9 月 16 日报道. http://edu.people.com.cn/GB/1053/3672381.html。

[3] Peter F. Brown, Stephan A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311.

[4] Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of ACL 2002, Philadelphia, PA: 295-302

[5] Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In Proceedings of ACL 2003, Sapporo, Japan: 160-167

[6] David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In Proceedings of ACL 2005, Ann Arbor, Michigan: 263-270

[7] Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In Proceedings of ACL 2001, Toulouse, France :523-530.

[8] Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In Proceedings of COLING/ACL 2006, Sydney, Australia:609-616

[9] Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In Proc. of ACL/HLT 2008:559-567

[10] Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In Proceedings of EMNLP 2007:61-72

[11] D. Xiong, M. Zhang, and H. Li. 2012. Modeling the translation of predicate-argument structure for smt. In Proc. of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation. In Proceedings of ACL 2012: 902-911

[12] 新科鉴字第 215 号，新疆民汉语机器翻译信息系统，新疆维吾尔自治区科学技术鉴定证书（98），1998 年 11 月。

[13] 哈力木拉提，吐尔根， 阿力甫，2005. 计算机汉、维辅助翻译软件研发. 第十届全国少数民族语言文字信息处理学术研讨会论文集，2005 年:112-115

[14] 艾山·毛力尼亚孜，谭勋，吐尔根·依布拉音，艾山·吾买尔. 2011. 维哈柯双语语料库加工系统词对齐技术的研究. 电脑知识与技术. 2011 年 7 卷(28 期): 6895-6897

[15] 热西旦·塔依，吐尔根·依布拉音. 2009. 汉文-维吾尔文双语语料库中基于词典译文的句子对齐方法研究. 新疆大学学报（自然科学版）. 2009 年 3 期 ：359-363

[16] 徐春，杨勇，董兴华. 汉维/维汉统计机器翻译中若干问题研究. 2011. 计算机工程与应用，2011 年 47 卷（35 期）: 150-154

[17] 任高举，吐尔根·伊布拉音，艾山·吾买尔. 2010. 统计机器翻译中汉维短语对抽取的研究. 新疆大学学报（自然科学版）. 2010 年 3 期 ：349-352

[18] 那顺乌日图，刘群，巴达玛敖德斯尔. 2001. 面向机器翻译的蒙古语生成. 自然语言理解与机器翻译，清华大学出版社，2001 年: 285-291

[19] 侯宏旭，刘群，那顺乌日图. 2007. 基于实例的汉蒙机器翻译. 中文信息学报. 2007 年，第 4 期: 65-72

[20] 胡冠龙，李淼. 2007. 基于逐次筛选法的多引擎汉民机器翻译系统. 民族语言文字信息技术研究，2007 年 2 月: 187-191

[21] 吉日木图. 2005. 基于模板的英蒙机器翻译系统的研究. 内蒙古大学硕士论文，2005 年: 1-55

[22] 百顺. 2008. 基于派生文法的日-蒙动词短语机器翻译研究. 中文信息学报. 2008 年第 22 卷(2 期): 47-52

[23] 娜步青. 2006. 基于统计的蒙汉机器翻译系统研究. 内蒙古农业大学学报（社会科学版），2006 年第 2 期

[24] 看卓才旦，金为勋，李延福，洛智华，朋毛扎西. 2006. 汉藏翻译系统中的动词处理研究。术语标准化

与信息技术，2006 年第 3 期 ：28-32

[25] 才藏太，华关加. 2005. 班智达汉藏公文翻译系统中基于二分法的句法分析方法研究. 中文信息学报，2005 年第 19 卷(第 6 期): 7-12

[26] 张国喜. 2004. 英藏命名实体在机器翻译系统的实现。青海师范大学学报（自然科学版），2004 年，第 3 期 ：26-28

[27] 才让加. 2011. 面向自然语言处理的大规模汉藏(藏汉)双语语料库构建技术研究. 中文信息学报。2011 年第 25 卷(6 期): 157-161

[28] 扎西加，高定国。 2011. 藏语语料库 TEI 标记规范探讨.中文信息学报. 2011 年第 25 卷(4 期): 66-70

[29] 多拉，扎西加，欧珠，大罗桑朗杰. 2007. 信息处理用藏文词类及标记集规范。第十一届全国民族语言文字信息学术研讨会，2007：441-452

[30] 杨攀，张建，李淼，乌达巴拉，雪艳. 2009. 汉蒙统计机器翻译中的形态学方法研究. 中文信息学报. 2009 年第 23 卷(1 期): 50-57

[31] 赵红梅，吕雅娟，贾国生，黄云，刘群. 2012. 第七届全国机器翻译研讨会机器翻译评测总结， 中文信息学报， 第 26 卷，第 1 期 ：22-30

[32] 张海波，刘凯，吕雅娟，华却才让，刘群. 2012. 一种通用的少数民族语言语种和编码识别方法，第四届全国少数民族青年自然语言信息处理学术研讨会, 2012 : 24-29

[33] 姜文斌，吴金星，长青，那顺乌日图，刘群，赵理莉. 2011. 蒙古语词法分析的有向图模型， 中文信息学报， 第 25 卷，第 5 期 ：94-100

[34] 麦热哈巴.艾力，姜文斌，王志洋，吐尔根.依不拉音，刘群. 2012. 维吾尔语词法分析的有向图模型. 软件学报，23(12):3115-3129

[35] 孙萌，华却才让，才智杰，姜文斌，吕雅娟，刘群. 2013. 基于判别式分类和重排序技术的藏文分词[J]，中文信息学报，已录用

[36] 王志洋，吕雅娟，刘群. 2011. 面向形态丰富语言的多粒度翻译融合， 中文信息学报， 第 4 期:75-81

[37] Zhiyang Wang, Yajuan Lü, Meng Sun, Qun Liu. 2013. Stem Translation with Affix-Based Rule Selection for Agglutinative Languages. In Proceedings of ACL 2013,Sofia, Bulgaria

[38] 付雷；黄瑾；何中军；吕雅娟；刘群. 2006. 一种融合了句型模板和统计机器翻译技术的翻译方法，中国，发明专利，专利号：ZL200610165532.6，授权日期：2009-09-23

[39] Zhiyang Wang , Yajuan Lü , Qun Liu. 2011. Multi-granularity Word Alignment and Decoding for Agglutinative Language Translation，Machine Translation Summit XIII (MT-summit XIII)，Xiamen, China: 360-368

[40] Kai Liu, Yajuan Lü, Wenbin Jiang and Qun Liu. 2013. Bilingually-Guided Monolingual Dependency Grammar Induction. In Proceedings of ACL 2013,Sofia, Bulgaria

作者简介：

吕雅娟： 中科院计算技术研究所、副研究员 lvyajuan@ict.ac.cn

刘　群： 中科院计算技术研究所、研究员

姜文斌： 中科院计算技术研究所、助理研究员